Term-Weighting for Summarization of Multi-party Spoken Dialogues

Term-Weighting for Summarization of Multi-Party Spoken Dialogues

Gabriel Murray and Steve Renals

University of Edinburgh, Edinburgh, Scotland,
gabriel.murray@ed.ac.uk, s.renals@ed.ac.uk,
http://www.cstr.ed.ac.uk

Abstract. This paper explores the issue of term-weighting in the genre of spontaneous, multi-party spoken dialogues, with the intent of using such term-weights in the creation of extractive meeting summaries. The field of text information retrieval has yielded many term-weighting techniques to import for our purposes; this paper implements and compares several of these, namely *tf.idf*, Residual IDF and *Gain*. We propose that term-weighting for multi-party dialogues can exploit patterns in word usage among participant speakers, and introduce the *su.idf* metric as one attempt to do so. Results for all metrics are reported on both manual and automatic speech recognition (ASR) transcripts, and on both the ICSI and AMI meeting corpora.

1 Introduction

The primary focus of this research is to create extractive summaries of meeting speech, in order to present users with concise and informative overviews of the content of meetings. Such extractive summaries, when incorporated into a meeting browser, can act as efficient tools for navigating meeting records as a whole. This paper focuses on one fundamental component of the extractive summarization pipeline: the way that terms are weighted within a given meeting, and the bearing that various term-weighting schemes have on extraction performance.

Choosing and implementing a term weighting method is often the first step in building an automatic summarization system. Though the unit of extraction may be the sentence or the dialogue act, those units need to be weighted by the importance of their constituent words. Popular text summarization techniques such as Maximal Marginal Relevance (MMR) and Latent Semantic Analysis (LSA) begin by representing sentences as vectors of term weights. There is a wide variety of term weighting schemes available, from simple binary weights of word presence/absence to more complex weighting schemes such as *tf.idf* and *tf.ridf*. Several of these are described in the following section.

A central question of this paper is whether term-weighting techniques developed for information retrieval (IR) and summarization tasks on text are well-suited for our domain of multiparty spontaneous spoken dialogues, or whether the patterns of word usage in such dialogues can be exploited in order to yield superior term-weighting for our task. To this end, we devise and implement a

novel term-weighting approach for multi-party speech called *su.idf*, based on differing word frequencies among speakers in a meeting. This metric is compared with 3 popular term-weighting schemes - *tf.idf*, *ridf* and *Gain* - and the metrics are evaluated via an extractive summarization task on both AMI and ICSI corpora.

2 Previous Term Weighting Work

Term weighting methods form an essential part of any IR system. Terms that characterize a given document well and discriminate the document from the remainder of the document collection should be weighted highly [1]. The most popular term weighting schemes have therefore combined *collection frequency* metrics with *term frequency* metrics. The latter component measures the term's prevalence in the document at hand while the former component analyzes the term usage across many documents.

The most common method of calculating collection frequency is called the *inverse document frequency* (IDF) [2]. The IDF for term t is given by

$$IDF(t) = -\log(\frac{D_w}{D})$$

or equivalently,

$$IDF(t) = \log D - \log D_w$$

where D is the total number of documents in the collection and D_w is the number of documents containing the term t. A term will therefore have a high IDF score if it is rare across the set of documents.

For the *term frequency* component, the simplest method is a binary term weight: 0 if the term is not present and 1 if it is. More commonly, the number of term occurrences in the document is used. Thus the term frequency TF is given by

$$TF(t,d) = \frac{N(t)}{\sum_{k=1}^{T} N(k)}$$

where $N(t)$ is the number of times the term t occurs in the given document and $\sum_{k=1}^{T} N_k$ is the total word count for the document, thereby normalizing the term count by document length.

The classic method for combining these components is simply *tf.idf* [1], wherein a term is scored highly if it occurs many times within a given document but rarely across the set of all documents. This term weighting scheme *tf.idf* increases our ability to discriminate between the documents in the collection. While there are variants to the TF and IDF components given above [1], the motivating intuitions are the same. Another example of combining these tree types of data (collection frequency, term frequency and document length)

is given by Robertson et al [3] and is called the Combined Weight. For a term t(i) and document d(j), the Combined Weight is described as:

$$CW(t,d) = \frac{IDF(t) \cdot TF(t,d) \cdot (K+1)}{K \cdot ((1-b) + (b \cdot (NDL(d)))) + TF(t,d)}$$

where K is a tuning constant regulating the impact of term frequency, b is a tuning constant regulating the impact of document length, and NDL is the normalized document length.

When relevance information is available, i.e. a subset of documents has been determined to be relevant to a user query, additional proven metrics are available for term relevance weighting and/or query expansion [3]. One example is the RSJ metric given in [4]:

$$RSJ(t,q) = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)}$$

where R is the number of documents known to be relevant to the query q and r is the number of relevant documents containing term t. The following variation is sometimes used instead, partly to avoid infinite weights under certain conditions:

$$RW(t,q) = \log(\frac{((r+0.5)(N-n-R+r+0.5))}{((n-r+0.5)(R-r+0.5))}$$

It is often the case, however, that there is little or no relevance information available when doing term weighting. Work by Croft and Harper [5] has shown that IDF is an approximation of the RSJ relevance weighting scheme when complete relevance information is unavailable. Robertson [6] further discusses the relationship between IDF and relevance weighting and places the IDF scheme on strong theoretical ground.

One extension of IDF called *ridf* [7] has proven effective for automatic summarization [8] and named entity recognition [9]. In *ridf*, the usual IDF component is substituted by the difference between the IDF of a term and its expected IDF according to the poisson model. The *ridf* score can be calculated by the formula

$$expIDF = -\log(1 - e^{(-f_w/D)})$$

$$ridf = IDF - expIDF$$

where f_w is the frequency of the word across all documents D.

Papineni [10] also provides an extension to IDF. Arguing that the IDF of a word is not synonymous with the *importance* of a word, but is rather an optimal weight for document self-retrieval, Papineni proposes a term-weighting metric *Gain* which is meant to measure importance or information gain of the term in the document:

$$Gain = \frac{D(t)}{D}(\frac{D(t)}{D} - 1 - \log\frac{D(t)}{D})$$

Very common and very rare words have low gain; this is in contrast with IDF, which will tend to give high scores to uncommon words. *ridf* also favors medium-frequency words [8]. As Papineni points out [10], the effective performance of metrics such as *ridf* and *Gain* seems to corroborate Luhn's observation that medium-frequency words have the optimal "resolving power" [11].

Mori et al [12] introduce a term weighting metric for automatic summarization called Information Gain Ratio (IGR). The underlying idea of IGR is that documents are clustered according to similarity, and further grouped into sub-clusters. If the information gain of a word increases after clusters are partitioned into sub-groups, then it can be said that the word contributes to that sub-cluster and should thus be rated highly.

Finally, Song et al [13] introduce a term weighting scheme for automatic summarization that is based on lexical chains. Building lexical chains in the manner of Barzilay [14], they weight chains according to how many word relations are in the chain, and weight each word in a chain according to how connected it is in the chain. On DUC 2001 data, they reported outperforming *tf* and *tf.idf* weighting schemes.

3 Term-Weighting for Meeting Speech

A common theme of most of the term-weighting metrics described in the previous section is that the distribution of words across a collection of documents is key to determining an ideal weight for the words. In general, words that are unique to a given document or cluster of documents should be weighted more highly than words that occur evenly throughout the entire document collection. For multiparty spoken dialogue, we have another potential source of variation in lexical usage: the speakers themselves. We introduce a new term weighting score for multi-party spoken dialogues by also considering how term usage varies across speakers in a given meeting. The intuition is that keywords will not be used by all speakers with the same frequency. Whereas IDF compares a given meeting to a set of all meetings, we can also compare a given speaker to a set of other speakers in the meeting. For each of the four speakers in a meeting, we calculate a surprisal score for each word that speaker uttered, which is the negative log probability of the term occuring amongst the other three speakers. The surprisal score for each word w uttered by speaker s is

$$surp(s, w) = -\log\left(\frac{\sum_{s' \neq s} tf(w, s')}{\sum_{r \neq s} N(r)}\right)$$

where *tf(w,s')* is the term frequency of word w for speaker *s'* and $N(r)$ is the total number of words spoken by each speaker r. For each term, we total its speaker surprisal scores and divide by the total number of speakers to find the overall surprisal score *surp(w)*. Thus the surprisal score for a word is given by

$$surp(w) = \frac{1}{S}\sum_{s} surp(s, w)$$

This surprisal score, the first component of the term-weighting metric, is then multiplied by $\frac{s(w)}{S}$, where $s(w)$ is the number of speakers who speak that word and S is the total number of speakers in the meeting. The third component of the metric is the inverse document frequency, or *idf*. The equation for *idf* is

$$idf(w) = -\log(\frac{D_w}{D})$$

where D is the total number of documents and D_w is the number of documents containing the term w. Putting these three components together, our term weighting metric is

$$su.idf = surp(w) \cdot \frac{s(w)}{S} \cdot \sqrt{idf}$$

One motivation for this novel term weighting scheme is that many important words in such meeting corpora are not necessarily rare across all documents, e.g. *cost*, *design* and *colour*. They are also not necessarily the most frequent content words in the meetings. They would therefore not score highly on either component of . Though we retain inverse document frequency for our new metric, the square root of *idf* is used to lower its overall influence within the metric, so that a term will not necessarily be weighted low if it is fairly common or weighted high simply because it is rare. Results on the development and test sets show a significant improvement by using the square root of *idf* rather than *idf* itself.

The hypothesis is that more informative words will be used with varying frequencies between the four meeting participants, whereas less informative words will be used fairly consistently by all. The component $\frac{s(w)}{S}$ is included for two reason. First, because individuals normally have idiosyncrasies in their speaking vocabularies, e.g. one meeting participant might use a type of filled pause not used by the others or otherwise frequently employ a word that is particular to their idiolect. And second, a word that is used by multiple speakers but with much different frequency should be more important than a word that is spoken by only one person.

There are several reasons for hypothesizing that use of informative words will vary between meeting participants. One is that meeting participants tend to have unique, specialized roles relevant to the discussion. In the AMI corpus, these roles are explicitly labelled, e.g. "marketing expert." With a given role comes a vocabulary associated with that role, e.g. "budget" and "cost" would be associated with a finance expert and "scroll" and "button" would be associated with an interface designer. Second, even when the roles are not so clearly defined, different participants have different areas of interest and different areas of expertise, and we expect that their vocabularies reflect these differences.

4 Experimental Setup

In addition to *tf.idf* and *su.idf*, we also implemented Residual IDF (*ridf*) and *Gain* for comparison. A hybrid approach combining the rankings of *tf.idf* and

su.idf was implemented in the hope that the two methods would be complementary, perhaps locating different types of informative terms. For all collection frequency measures, we used a collection of documents from the AMI, ICSI, Broadcast News and MICASE corpora. All term-weighting methods were run on both manual and ASR transcripts.

4.1 Data Description

We tested our term-weighting methods on the AMI and ICSI meeting corpora, which differ from one another in several important ways. The AMI meeting corpus [15] is a corpus of both scenario and non-scenario meetings, though for these experiments we used only scenario meetings. In these scenario meetings, four participants take part in each meeting and play roles within a fictional company. The scenario given to them is that they are part of a company called Real Reactions, which designs remote controls. Their assignment is to design and market a new remote control, and the members play the roles of project manager (the meeting leader), industrial designer, user-interface designer, and marketing expert. Through a series of four meetings, the team must bring the product from inception to market. The participants are also given real-time information from the company during the meetings, such as information about user preferences and design studies, as well as updates about the time remaining in each meeting. While the scenario given to them is artificial, the speech and the actions are completely spontaneous and natural.

The AMI test set consists of 19 meetings, or 4 sequences of 4 meetings each and 1 sequence of 3 meetings.

The second corpus used herein is the ICSI meeting corpus [16], a corpus of 75 natural, i.e. non-scenario, meetings, approximately one hour each in length. The ICSI test set consists of 6 meetings.

ASR for both corpora was kindly provided by the AMI-ASR group. The word-error rate (WER) for the AMI corpus is 43% while the WER for the ICSI corpus is 29.5%.

For both corpora, multiple human annotations were carried out for evaluation purposes. A human-authored abstract is created for each meeting, summarizing the most important aspects of the meeting in terms of decision, actions and goals of the meeting. Multiple human annotators then work through the meeting transcript and link dialogue acts to sentences in the human abstract when they find that a given dialogue act supports an abstract sentence. The result is a many-to-many mapping between dialogue acts and sentences in the abstract, so that a given dialogue act can be linked to more than one abstract sentence, and vice-verse.

4.2 Evaluation Protocol

For our evaluation, each term-weighting approach was used to create a brief summary of each test set meeting, and the resulting summaries were then evaluated. In each case we summed term-scores over dialogue acts to create scores for the

dialogue acts, which are the summary extraction unit. Dialogue acts are ranked from most informative to least informative, and are extracted until a length of 700 words is reached. These summaries are then evaluated using the *weighted precision* metric originally introduced by Murray et al [17]. This metric is based on the multiple human annotations of dialogue act importance described above. Because each annotator creates a many-to-many mapping between dialogue acts and sentences within the human abstract, we can score each summary dialogue act according to how often each annotator linked it, and score the summary overall based on the constitutent dialogue act scores.

5 Results

The following sections detail the results on both the AMI and ICSI corpora.

5.1 AMI Results

Meet	sidf	sasr	tfidf	tfasr	com	comasr	ridf	ridfasr	gain	gainasr
ES2004a	0.50	0.51	0.50	0.59	0.55	0.55	0.59	0.64	0.63	0.63
ES2004b	0.59	0.67	0.58	0.55	0.59	0.60	0.67	0.65	0.64	0.69
ES2004c	0.66	0.63	0.69	0.64	0.76	0.67	0.76	0.71	0.59	0.67
ES2004d	0.69	0.75	0.85	0.77	0.99	0.78	0.77	0.79	0.78	0.85
ES2014a	0.67	0.70	0.68	0.71	0.67	0.71	0.70	0.76	0.65	0.73
ES2014b	0.76	0.81	0.74	0.70	0.86	0.72	0.79	0.75	0.77	0.83
ES2014c	0.74	0.78	0.69	0.67	0.88	0.77	0.83	0.80	0.69	0.71
ES2014d	0.51	0.40	0.44	0.40	0.48	0.44	0.43	0.36	0.33	0.43
IS1009a	0.85	0.73	0.68	0.72	0.69	0.73	0.74	0.78	0.74	0.73
IS1009b	0.65	0.83	0.50	0.68	0.57	0.70	0.65	0.73	0.57	0.78
IS1009c	0.50	0.52	0.34	0.36	0.46	0.42	0.44	0.45	0.44	0.56
IS1009d	0.74	0.60	0.73	0.58	0.81	0.71	0.75	0.69	0.58	0.50
TS3003a	0.53	0.50	0.48	0.48	0.54	0.52	0.57	0.60	0.63	0.61
TS3003b	0.63	0.73	0.64	0.59	0.57	0.55	0.68	0.67	0.70	0.72
TS3003c	0.89	0.93	0.89	0.87	0.86	0.90	0.80	0.79	0.80	0.92
TS3003d	0.46	0.54	0.41	0.51	0.46	0.54	0.59	0.56	0.63	0.63
TS3007a	0.37	0.54	0.35	0.54	0.37	0.52	0.50	0.57	0.51	0.57
TS3007b	0.62	0.61	0.57	0.54	0.66	0.56	0.67	0.62	0.59	0.59
TS3007c	0.70	0.64	0.61	0.48	0.64	0.60	0.55	0.57	0.60	0.73
AVERAGE	**0.63**	**0.65**	**0.60**	**0.60**	**0.65**	**0.63**	**0.66**	**0.66**	**0.62**	**0.68**

Table 1. Weighted Precision Results for AMI Test Set Meetings

sidf=*su.idf* on manual, **sasr**=*su.idf* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **comasr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=residual IDF on manual, **ridfasr**=residual IDF on ASR, **gain**=*Gain* on manual, **gainasr**=*Gain* on ASR

On manual transcripts, the best approaches were *su.idf*, the hybrid approach combining *su.idf* and *tf.idf*, and *ridf*, all of which were significantly better than *tf.idf* ($p>0.95$) and not significantly different from one another, according to paired t-tests.

On ASR transcripts, *Gain* performed much better than it had on manual transcripts, with higher weighted precision results than the other approaches.

su.idf also performed better on ASR, with its weighted precision increasing from 0.63 to 0.65. The hybrid approach slipped 2 points, while the *tf.idf* weighted precision scores stayed much the same. *Gain*, *su.idf* and *ridf* all performed significantly better than *tf.idf*. Table 1 gives results on both manual and ASR.

Meet	Summ-WER	NonSumm-WER
ES2004a	47.1	56.0
ES2004b	35.5	45.9
ES2004c	34.8	48.1
ES2004d	43.6	54.8
ES2014a	43.5	56.9
ES2014b	37.4	53.9
ES2014c	43.9	54.5
ES2014d	39.6	53.6
IS1009a	42.6	50.0
IS1009b	43.9	48.5
IS1009c	59.2	57.6
IS1009d	46.3	46.5
TS3003a	26.7	45.2
TS3003b	25.2	30.3
TS3003c	22.7	34.8
TS3003d	27.9	38.2
TS3007a	33.6	44.3
TS3007b	27.1	38.3
TS3007c	31.8	42.7
AVERAGE	**37.49**	**47.37**

Table 2. Word Error Rates for Extracted (**Summ-WER**) and Non-Extracted Portions (**NonSumm-WER**) of Meetings

It was particularly surprising that some of the term-weighting approaches performed better on ASR than on manual transcripts. Previous research [18, 19] has shown that informative portions of speech data tend to have lower word-error rates, but it is nonetheless unexpected that weighted precision would actually *improve* on errorful ASR transcripts. *Gain* and *su.idf* were particularly resilient to the errorful transcripts on this test set. Table 2 shows the word-error rates for the extracted and non-extracted portions of meetings using the *su.idf* summarizer. The WER for the extracted portions is nearly 10 points lower than for the non-extracted portions of meetings, at 37.49% versus 47.37%. The WER for the corpus as a whole is around 43.0%.

To get a better idea of how *su.idf* and *tf.idf* differ in the way they score and rank terms, and in particular why the performance gap increases on ASR, we plotted term-score against term-rank for both metrics on one of the AMI test set meetings, TS3003b. On manual transcripts, performance according to weighted precision was comparable for this meeting. However, on ASR transcripts weighted precision for *su.idf* increased by 10 points while the scores for *tf.idf* decreased by 5 points. As Figure 1 shows, the relationship between term-score and term-rank varies greatly depending on the metric. *tf.idf* tends to score only a few words highly, so that there is a sudden drop-off in scores for words that are ranked only slightly lower. In contrast, *su.idf* tends to score a larger number of words highly and the descent of scores is less steep as the rank decreases.

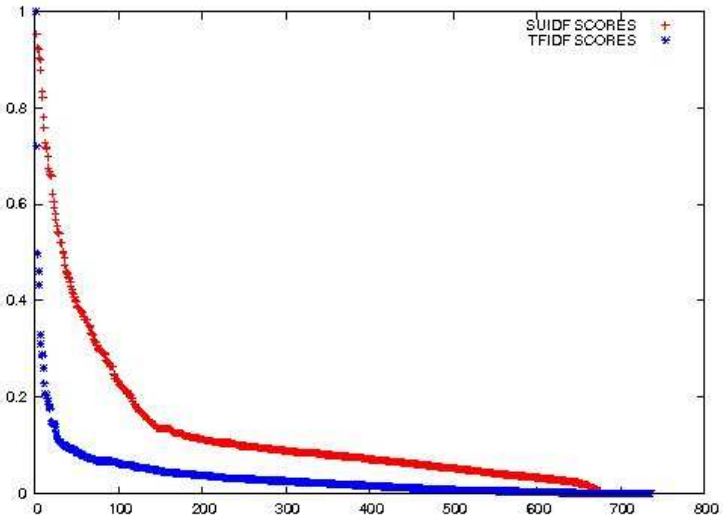

Fig. 1. *Term Rank Plotted Against Term Score, ASR Transcripts*

This trend is found across meetings, and the difference between the approaches is particularly pronounced on ASR.

5.2 ICSI Results

On both manual and ASR transcripts there were fewer differences between term-weighting approaches than were found on the AMI test set. On manual transcripts, the highest scoring approaches were *Gain* and the hybrid of *su.idf* and *tf.idf*; however, there were no significant differences between approaches as a whole. As can be seen in Table 3, the weighted precision scores in general are much lower than on the AMI meetings.

Meet	sidf	sasr	tfidf	tasr	com	comasr	ridf	ridfasr	gain	gainasr
Bed004	0.28	0.35	0.32	0.37	0.33	0.41	0.33	0.35	0.35	0.38
Bed009	0.53	0.45	0.44	0.38	0.45	0.42	0.38	0.38	0.39	0.39
Bed016	0.38	0.47	0.52	0.53	0.46	0.56	0.59	0.62	0.50	0.46
Bmr005	0.44	0.44	0.44	0.49	0.52	0.44	0.53	0.54	0.53	0.55
Bmr019	0.37	0.33	0.25	0.31	0.34	0.41	0.30	0.32	0.35	0.40
Bro018	0.36	0.36	0.39	0.32	0.41	0.36	0.36	0.32	0.39	0.29
AVERAGE	**0.39**	**0.40**	**0.39**	**0.40**	**0.42**	**0.43**	**0.42**	**0.42**	**0.42**	**0.41**

Table 3. Weighted Precision Results for ICSI Test Set Meetings

sidf=*su.idf* on manual, **sasr**=*su.idf* on ASR, **tfidf**=*tf.idf* on manual, **tfasr**=*tf.idf* on ASR, **com**=combined *su.idf* and *tf.idf* on manual, **comasr**=combined *su.idf* and *tf.idf* on ASR, **ridf**=residual IDF on manual, **ridfasr**=residual IDF on ASR, **gain**=*Gain* on manual, **gainasr**=*Gain* on ASR

On ASR, the highest scoring method was the hybrid approach, followed by *ridf*. There were again no significant differences between the various methods. Interestingly, however, all approaches tended to do better on ASR than on manual transcripts, as evidenced previously on the AMI test set above. Surprisingly, the only approach that showed decreasing weighted precision scores on ASR was *Gain*, which slipped by a point. This is in contrast to the AMI results, where *Gain* did significantly better on ASR transcripts than on manual.

6 Discussion

There are several interesting aspects of the results reported above. Perhaps the most surprising is that some of the metrics, especially *su.idf* and *Gain*, are particularly resilient to ASR errors, and we found a general trend that weighted precision actually increased on ASR.

We also found that most of our metrics easily outperformed the classic *tf.idf* term-weighting scheme, with *su.idf*, the hybrid approach and *ridf* consistently performing the best. While *su.idf* outperformed *tf.idf* on the AMI corpus meetings, there was no statistical difference between the two approaches on the ICSI meetings. However, it was still advantageous to calculate *su.idf* on those meetings, as the hybrid approach was superior. Part of the reason for the difference in performance of those two metrics on AMI versus ICSI meetings may be due to the structure and set-up of the meetings themselves. As described above, the AMI meetings are scenario meetings with well-defined roles such as *project manager* and *marketing expert*, whilst roles in the ICSI corpus are much less clearly defined. Because roles are associated with certain vocabularies (e.g. the marketing expert being more likely to say "trend" or "survey" than the others), perhaps it would be expected that *su.idf* would perform better on those meetings than on meetings where roles are more opaque and the structure of the meetings is more loosely defined. Having said that, there were no significant differences between *any* of the term-weighting approaches on the ICSI meetings, and the results on a smaller test-set may simply be less reliable.

One clear result is that *tf.idf* is not as sensitive to term importance as the other metrics. It seems telling then that it is also the only metric that weights a term highly for occurring frequently within the given document. It is perhaps too blunt, favoring a few terms by scoring them highly and scoring the others dramatically lower, leading to a severely limited view of importance within the meeting. A strength of *su.idf* is that a term need not be very frequent within a document nor very rare across documents in order to receive a high score.

Our evaluation has relied on weighted precision of summaries that were created using each term-weighting scheme. We currently limit the evaluation to precision because the summaries are very brief and subsequently all recall scores are quite small. In the future we may wish to expand our evaluation to weighted precision, recall and f-measure, perhaps using longer automatic summaries. The weighted precision metric also, as currently formulated, does not have a theoretical maximum due to the fact that annotators may link each dialogue act as

many times as they wish. One solution would be to use only the number of annotators who link each dialogue act, rather than the number of links they give to each dialogue act, thus providing a maximum score across summaries. However, doing so would cause us to lose a substantial amount of information in the form of annotator link counts.

7 Future Work

While exploiting differences in term usage among speakers has been promising, we believe there are additional speech-based features to exploit for term-weighting. One example is that informative terms used in meeting speech should tend to cluster into portions of the meeting roughly correlating to topic structure, whereas less informative words will be spread throughout the meeting. In addition, measures of prosodic prominence such as energy and F0 variance may be informative for locating more important words within the meeting.

8 Conclusion

We have presented an evaluation of term-weighting metrics for spontaneous, multi-party spoken dialogues. Three of the metrics, *tf.idf*, *ridf* and *Gain*, were imported from text IR to test for suitability with our data. A novel approach called *su.idf* was implemented, relying on the differing patterns of word usage among meeting participants. It was found to perform very competitively, both on its own and as part of a hybrid approach using combined rankings with *tf.idf*. In addition to the encouraging results for *su.idf*, we have provided evidence that *ridf* and *Gain* outperform *tf.idf* on our speech data.

9 Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for speech recognition output. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-245)

References

1. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24** (1988) 513–523
2. Jones, K.S.: A statistical interpretation of term specificity and its application in retrievall. Journal of Documentation **28** (1972) 11–21
3. Robertson, S., Jones, K.S.: Simple, proven approaches to text retrieval. University of Cambridge Computer Laboratory Technical Report TR-356 **356** (1994)
4. Robertson, S., Jones, K.S.: Relevance weighting of search terms. Journal of the American Society for Information Science **35** (1976) 129–146
5. Croft, W., Harper, D.: Using probabilistic models of information retrieval without relevance information. Journal of Documentation **35** (1979) 285–295
6. Robertson, S.: Understanding inverse document frequency: On theoretical arguments for idf. Journal of Documentation **60** (2004) 503–520
7. Church, K., Gale, W.: Inverse document frequency IDF: A measure of deviation from poisson. In: Proc. of the Third Workshop on Very Large Corpora. (1995) 121–130
8. Orasan, C., Pekar, V., Hasler, L.: A comparison of summarisation methods based on term specificity estimation. In: Proc. of LREC 2004, Lisbon, Portugal. (2007) 1037–041
9. Rennie, J., Jaakkola, T.: Using term informativeness for named entity recognition. In: Proc. of SIGIR 2005, Salvador, Brazil. (2005) 353–360
10. Papineni, K.: Why inverse document frequency? In: Proc. of NAACL 2001. (2001) 1–8
11. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, NY, NY, USA (1983)
12. Mori, T.: Information gain ratio as term weight: The case of summarization of ir results. In: Proc. of COLING 2002, Taipei, Taiwan. (2002) 688–694
13. Song, Y., Han, K., Rim, H.: A term weighting method based on lexical chain for automatic summarization. In: Proc. of CICLing 2004, Seoul, Korea. (2004) 636–639
14. Barzilay, R., Elhadad, M.: Using lexical chains for summarisation. In: Proc. of ACL 1997, Madrid, Spain. (1997) 10–18
15. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Proc. of MLMI 2005, Edinburgh, UK. (2005) 28–39
16. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C.: The ICSI meeting corpus. In: Proc. of IEEE ICASSP 2003, Hong Kong, China. (2003) 364–367
17. Murray, G., Renals, S., Moore, J., Carletta, J.: Incorporating speaker and discourse features into speech summarization. In: Proc. of the HLT-NAACL 2006, New York City, USA. (2006) 367–374
18. Valenza, R., Robinson, T., Hickey, M., Tucker, R.: Summarization of spoken audio through information extraction. In: Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK. (1999) 111–116
19. Murray, G., Renals, S., Carletta, J.: Extractive summarization of meeting recordings. In: Proc. of Interspeech 2005, Lisbon, Portugal. (2005) 593–596